Correlated z-values and the accuracy of large-scale statistical estimates

Bradley Efron*†

Abstract

We consider large-scale studies in which there are hundreds or thousands of correlated cases to investigate, each represented by its own normal variate, typically a z-value. A familiar example is provided by a microarray experiment comparing healthy with sick subjects' expression levels for thousands of genes. This paper concerns the accuracy of summary statistics for the collection of normal variates, such as their empirical cdf or a false discovery rate statistic. It seems like we must estimate an N by N correlation matrix, N the number of cases, but our main result shows that this is not necessary: good accuracy approximations can be based on the root mean square correlation over all $N \cdot (N-1)/2$ pairs, a quantity often easily estimated. A second result shows that z-values closely follow normal distributions even under non-null conditions, supporting application of the main theorem. Practical application of the theory is illustrated for a large leukemia microarray study.

Key words: rms correlation, non-null z-values, correlation penalty, Mehler's identity, empirical process, acceleration

1 Introduction

Modern scientific studies routinely produce data on thousands of related situations. A familiar example is a microarray experiment in which thousands of genes are being investigated for possible disease involvement. Each gene might produce a z-value, say z_i, for the ith gene, by definition a test statistic theoretically having a standard normal distribution

$$H_0 : z_i \sim \mathcal{N}(0,1) \tag{1.1}$$

under the null hypothesis H_0 of no disease involvement. A great deal of the current literature was developed under the assumption of independence among the z_i's. This can be grossly unrealistic in practice, as discussed in Owen (2005) and Efron (2007a), among others. This

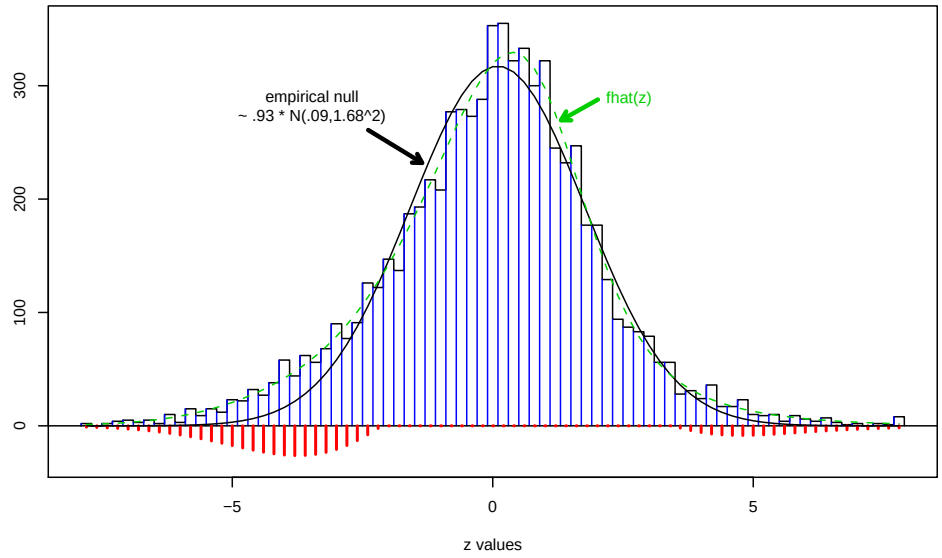

Figure 1: Histogram of z-values for $N = 7128$ genes, leukemia study, Golub et al. (1999). *Dashed curve* $\hat{f}(x)$, smooth fit to histogram; *solid curve* "empirical null", normal density fit from central 50% of histogram, is much wider than theoretical $\mathcal{N}(0,1)$ null distribution. Small red bars plotted negatively discussed in Section 4.

paper concerns the accuracy of summary statistics of the z_i's, for example, their empirical cdf (cumulative distribution function), under conditions of substantial correlation.

Figure 1 concerns a leukemia microarray study by Golub et al. (1999) that we will use for motivation and illustration. Two forms of leukemia are being examined for possible genetic differences: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In the version of the data discussed here there are $n_1 = 47$ ALL patients and $n_2 = 25$ AML patients, with expression levels on the same $N = 7128$ genes measured on each patient.

A two-sample t-statistic t_i comparing AML with ALL expression levels was computed for each gene and converted to a z-value,

$$z_i = \Phi^{-1}\left(F_{70}(t_i)\right) \qquad i = 1, 2, \ldots, N, \tag{1.2}$$

where Φ and F_{70} are the cumulative distribution functions for a standard normal and a Student-t distribution with 70 degrees of freedom. Figure 1 shows a histogram of the z_i's, which turns out to be much wider than (1.1) suggests: its central spread is estimated to be $\hat{\sigma}_0 = 1.68$ rather than 1, as discussed in Section 3.

Here is an example of the results to be derived in Sections 2 through 4. Let $\hat{F}(x)$ be the right-sided cdf ("survival curve") of the z-values,

$$\hat{F}(x) = \#\{z_i > x\}/N. \tag{1.3}$$

*Department of Statistics, Stanford University

†This work was supported in part by NIH grant 8R01 EB002784 and NSF grant DMS0505673.

x :	1	2	3	4	5
$\widehat{\mathrm{sd}}$:	**.017**	**.022**	**.0101**	**.0040**	**.0019**
$\widehat{\mathrm{sd}}_0$:	.005	.004	.0027	.0018	.0012
$\widehat{\mathrm{sd}}_{\mathrm{perm}}$:	.021	.001	.0014	.0001	.0000
$\hat{F}(x)$:	.29	.13	.057	.025	.010
$\widehat{\mathrm{Fdr}}(x)$:	.94	.92	.71	.38	.15

Table 1: Estimates of standard deviation for right-sided cdf $\hat{F}(x)$ (1.3); $\widehat{\mathrm{sd}}$ square root of formula (1.4); $\widehat{\mathrm{sd}}_0$ square root of first term in (1.4); $\widehat{\mathrm{sd}}_{\mathrm{perm}}$ permutation standard deviation. Accuracy of False Discovery Rate estimate $\widehat{\mathrm{Fdr}}(x)$ discussed in Section 4.

Then a good approximation for the variance of $\hat{F}(x)$ is

$$\mathrm{Var}\left\{\hat{F}(x)\right\} \doteq \left\{\frac{\hat{F}(x)\left(1-\hat{F}(x)\right)}{N}\right\} + \left\{\frac{\hat{\sigma}_0^2\hat{\alpha}\hat{f}^{(1)}(x)}{\sqrt{2}}\right\}^2. \tag{1.4}$$

The first term in (1.4) is the usual binomial variance, while the second term is a *correlation penalty* accounting for dependence between the z_i's. The quantities occurring in the correlation penalty are

- $\hat{\sigma}_0$, the estimate of central spread (1.68 above);
- $\hat{\alpha}$, an estimate of the root-mean-square of the correlations between the $N(N-1)/2$ pairs of z_i's (equaling about .11 for the leukemia data, as calculated from the simple formula in Section 3);
- $\hat{f}^{(1)}(x)$, the first derivative of a smooth fit to the z-value histogram (estimated by a Poisson spline regression in Figure 1).

The row marked $\widehat{\mathrm{sd}}$ in Table 1 is the square root of formula (1.4) applied to the leukemia data. $\hat{F}(4) = .025$ is seen to have $\widehat{\mathrm{sd}} = .0040$, more than double $\widehat{\mathrm{sd}}_0 = .0018$, the binomial standard deviation obtained by ignoring the second term in (1.4). The permutation standard deviation, obtained from repeated permutations of the 72 patients, is only .0001 at $x = 4$. Permutation methods, which preserve within-microarray correlations, have been advocated for large-scale hypothesis testing (see Westfall and Young, 1993; Dudoit, Shaffer and Boldrick, 2003, Sect. 2.6), but they are inappropriate for the accuracy considerations of this paper.

Formula (1.4), and more ambitious versions that include covariances across different values of x, are derived in Section 2 and Section 3: an exact expression is derived first, followed by a series of simplifying approximations and techniques for their estimation. The basic results are extended to provide more general accuracy estimates in Section 4: comparing, for example, the variability of local versus tail-area false discovery rates.

All of our results depend on the assumption that the z_i's are normal with possibly different means and variances,

$$z_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \qquad i = 1, 2, \ldots, N. \tag{1.5}$$

There is no requirement that they be "z-values" in the hypothesis-testing sense of (1.1), and in fact this paper is more concerned with estimation than testing. However, z-values are ubiquitous in large-scale applications, and not only in the two-sample setting of the leukemia study. Section 5 concerns the *non-null distribution of z-values*. A theorem is derived justifying (1.5) as a good approximation, allowing results like (1.4) to be applied to the leukemia study z-values. Section 6 and Section 7 close with remarks and a brief summary.

The statistics microarray literature has shown considerable interest in the effects of large-scale correlation, some good references being Dudoit, van der Laan and Pollard (2004), Owen (2005), Qiu, Klebanov and Yakovlev (2005b), Qiu, Brooks, Klebanov and Yakovlev (2005a) and Desai, Deller and McCormick (2009). Efron (2007a) used a z-value setting to examine the effects of correlation on false discovery rate analysis; that paper's Section 2 theorem is a null hypothesis version of the general result developed here. A useful extension along different lines appears in Schwartzman and Lin (2009).

Clarke and Hall's (2009) asymptotic calculations support the use of the independence standard deviation $\widehat{\mathrm{sd}}_0$ in Table 1, even in the face of correlation. The situations they consider are low-correlation by the standard here, with the root-mean-square value $\hat{\alpha}$ of (1.4) approaching zero (from their assumption (3.2)). Since $\hat{\alpha}$ is often easy to estimate, formulas such as (1.4) provide a quantitative check on the use of $\widehat{\mathrm{sd}}_0$.

2 The distribution of correlated normal variates

Given N correlated normal variates $z_1, z_2, \ldots, z_N$, with possibly different means and standard deviations, let $\hat{F}(x)$ denote their right-sided empirical cdf[1]

$$\hat{F}(x) = \#\{z_i \geq x\}/N, \qquad \text{for } -\infty < x < \infty. \tag{2.1}$$

This section presents tractable formulas for the mean and covariance of the process $\{\hat{F}(x), -\infty < x < \infty\}$, and a simpler approximation that we will see is nicely suited for applications.

Rather than work directly with cdfs, it will be easier, and in a sense more basic, to first derive results for a discretized version of the empirical *density* of the z_i values. We partition the range $\mathcal{Z}$ into K bins $\mathcal{Z}_k$,

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k, \tag{2.2}$$

each bin being of width Δ. Let x_k indicate the midpoint of $\mathcal{Z}_k$, and y_k the number of z_i's in $\mathcal{Z}_k$,

$$y_k = \#\{z_i \in \mathcal{Z}_k\} \qquad k = 1, 2, \ldots, K. \tag{2.3}$$

[1] It is convenient for the applications of Section 4 to deal with right-sided cdfs or *survival curves* instead of the usual left-sided ones in (1.2), and we will use this definition in what follows.

We will derive expressions for the mean and covariance of the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)'$. In effect, $\boldsymbol{y}$ is the order statistic of $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)'$, becoming exactly that as the bin width $\Delta \to 0$. (In which case the y_k values go to 1 or 0, with the non-zero bin x_k values indicating the locations of the ordered z_i's, assuming no ties.) Familiar statistical applications, of the type described in Section 4, depend on $\boldsymbol{z}$ only through $\boldsymbol{y}$.

Suppose that the z_i's are divided into a finite number of classes, with members of the cth class $\mathcal{C}_c$ having mean μ_c and standard deviation σ_c,

$$z_i \sim \mathcal{N}(\mu_c, \sigma_c^2) \qquad \text{for } z_i \in \mathcal{C}_c. \tag{2.4}$$

Let N_c be the number of members of $\mathcal{C}_c$, with p_c the proportion

$$N_c = \#\{\mathcal{C}_c\} \quad \text{and} \quad p_c = N_c/N \tag{2.5}$$

so $\sum_c N_c = N$ and $\sum_c p_c = 1$. The use of model (2.4) for z-values is supported by the results of Section 5.

If $\boldsymbol{x}$ is the K-vector of bin midpoints, let $x_{kc} = (x_k - \mu_c)/\sigma_c$ and

$$\boldsymbol{x}_c = (\boldsymbol{x} - \mu_c)/\sigma_c = (\ldots, x_{kc}, \ldots)'. \tag{2.6}$$

Likewise, for any real-valued function $h(x)$ we define $\boldsymbol{h}_c$ to be the K-vector of function values

$$\boldsymbol{h}_c = (\ldots, h(x_{kc}), \ldots)', \tag{2.7}$$

also denoted by $h(\boldsymbol{x}_c)$ in what follows.

It is easy to calculate the expectation of the count vector $\boldsymbol{y}$ under the multi-class model (2.4)–(2.5). Let π_{kc} equal the probability that z_i from class $\mathcal{C}_c$ falls into the kth bin,

$$\pi_{kc} = \text{Prob}_c\{z_i \in \mathcal{Z}_k\} \doteq \Delta\varphi(x_{kc})/\sigma_c. \tag{2.8}$$

Here $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, the standard normal density. The approximation $\pi_{kc} \doteq \Delta\varphi(x_{kc})/\sigma_c$ from (2.4) becomes arbitrarily accurate for Δ sufficiently small, and we will take it as exact in what follows. Then

$$E\{\boldsymbol{y}\} = N\sum_c p_c\boldsymbol{\pi}_c = N\Delta\sum_c p_c\varphi(\boldsymbol{x}_c) = N\Delta\sum_c p_c\boldsymbol{\varphi}_c. \tag{2.9}$$

The $K \times K$ covariance matrix of the count vector $\boldsymbol{y}$ depends on the $N \times N$ correlation matrix of $\boldsymbol{z}$, but in a reasonably simple way discussed next. Two important definitions are needed to state the first result: there are $M = N(N-1)/2$ correlations $\rho_{ii'}$ between pairs $(z_i, z_{i'})$ of members of $\boldsymbol{z}$, and we denote by "$g(\rho)$" the distribution putting weight $1/M$ on each $\rho_{ii'}$. Also, for $\varphi_\rho(u, v)$ the bivariate normal density having zero means, unit standard deviations, and correlation ρ, we define

$$\lambda_\rho(u, v) = \frac{\varphi_\rho(u, v)}{\varphi(u)\varphi(v)} - 1 = (1-\rho^2)^{-\frac{1}{2}} \exp\left\{\frac{2\rho uv - \rho^2(u^2+v^2)}{2(1-\rho^2)}\right\} - 1 \tag{2.10}$$

and

$$\lambda(u,v) = \int_{-1}^{1} \lambda_\rho(u,v) g(\rho) d\rho \tag{2.11}$$

(the integral notation being shorthand for summing over M discrete points).

Lemma 1. *Under the multi-class model* (2.4)–(2.5)*, the covariance of the count vector* $\boldsymbol{y}$ (2.3) *has two components,*

$$\mathbf{cov}\,(\boldsymbol{y}) = \mathbf{cov}_0 + \mathbf{cov}_1 \tag{2.12}$$

where

$$\mathbf{cov}_0 = N \sum_c p_c \left\{ \mathrm{diag}(\boldsymbol{\pi}_c) - \boldsymbol{\pi}_c \boldsymbol{\pi}_c' \right\} \tag{2.13}$$

and

$$\begin{aligned} \mathbf{cov}_1 = N^2 \sum_c \sum_d p_c p_d \,\mathrm{diag}(\boldsymbol{\pi}_c) \boldsymbol{\lambda}_{cd} \,\mathrm{diag}(\boldsymbol{\pi}_d) \\ - N \sum_c p_c \,\mathrm{diag}(\boldsymbol{\pi}_c) \boldsymbol{\lambda}_{cc} \,\mathrm{diag}(\boldsymbol{\pi}_c). \end{aligned} \tag{2.14}$$

Here $\mathrm{diag}(\boldsymbol{\pi}_c)$ is the $K \times K$ diagonal matrix having diagonal elements π_{kc}, similarly $\mathrm{diag}(\boldsymbol{\pi}_d)$, while $\boldsymbol{\lambda}_{cd}$ is the $K \times K$ matrix with klth element $\lambda(x_{kc}, x_{ld})$; the summations are over all classes.

Note. Equation (2.14) assumes that the correlation distribution $g(\rho)$ is the same across all classes $\mathcal{C}_c$. The proof of Lemma 1, which is similar to that for the simpler situation of Efron (2007a), appears in Remark C of Section 6.

The $\mathbf{cov}_0$ term in (2.12)–(2.13) is the sum of the multinomial covariance matrices that would apply if the z_i's were mutually independent with fixed numbers drawn from each class; $\mathbf{cov}_1$ is a penalty for correlation, almost always increasing $\mathbf{cov}(\boldsymbol{y})$. The N^2 factor in (2.14) makes the correlation penalty more severe as N increases, assuming $g(\rho)$ stays the same.

Expression (2.14) for the correlation penalty can be considerably simplified. *Mehler's identity* for $\lambda_\rho(u,v)$ (2.10) is

$$\lambda_\rho(u,v) = \sum_{j \geq 1} \frac{\rho^j}{j!} h_j(u) h_j(v) \tag{2.15}$$

where h_j is the jth Hermite polynomial. (See Lancaster, 1958 for an enlightening discussion of (2.15), also known as the "tetrachoric series", and its connections to the singular value decomposition, canonical correlation, Pearson's coefficient of contingency, and correspondence analysis.) Denoting the jth moment of the correlation distribution $g(\rho)$ by α_j,

$$\alpha_j = \int_{-1}^{1} \rho^j g(\rho) d\rho, \tag{2.16}$$

(2.11) becomes

$$\lambda(u,v) = \sum_{j\geq 1} \frac{\alpha_j}{j!} h_j(u) h_j(v) \tag{2.17}$$

so $\boldsymbol{\lambda}_{cd}$ in (2.14) can be written in outer product notation as

$$\boldsymbol{\lambda}_{cd} = \sum_{j\geq 1} \frac{\alpha_j}{j!} h_j(\boldsymbol{x}_c) h_j(\boldsymbol{x}_d)'. \tag{2.18}$$

Making use of (2.8), taken as exact,

$$\begin{aligned} \text{diag}(\boldsymbol{\pi}_c) h_j(\boldsymbol{x}_c) &= N\Delta \text{diag}\left(\varphi(\boldsymbol{x}_c)\right) h_j(\boldsymbol{x}_c)/\sigma_c \\ &= (-1)^j N\Delta \cdot \boldsymbol{\varphi}_c^{(j)}/\sigma_c \end{aligned} \tag{2.19}$$

where $\boldsymbol{\varphi}_c^{(j)}$ indicates the jth derivative of $\varphi(u)$ evaluated at each component of $\boldsymbol{x}_c$ (using $\varphi^{(j)}(u) = (-1)^j \varphi(u) h_j(u)$).

Rearranging (2.14) then gives a simplified formula.

Lemma 2. *Defining*

$$\bar{\boldsymbol{\phi}}^{(j)} \equiv \sum_c p_c \boldsymbol{\varphi}_c^{(j)}/\sigma_c, \tag{2.20}$$

(2.14) *for the correlation penalty becomes*

$$\mathbf{cov}_1 = N^2\Delta^2 \left\{ \sum_{j\geq 1} \frac{\alpha_j}{j!} \bar{\boldsymbol{\phi}}^{(j)} \bar{\boldsymbol{\phi}}^{(j)'} - \frac{1}{N} \sum_{j\geq 1} \frac{\alpha_j}{j!} \left(\sum_c p_c \boldsymbol{\varphi}_c^{(j)} \boldsymbol{\varphi}_c^{(j)'}/\sigma_c^2 \right) \right\}. \tag{2.21}$$

A convenient approximation to $\mathbf{cov}_1$ is based on three reductions of (2.21):

- The second term in (2.21) is neglible for large N.
- Common standardization methods for large-scale data sets often make α_1, the expectation of $g(\rho)$, exactly or nearly zero, as illustrated in Section 3 for the leukemia data; see Section 3 of Efron (2009).
- This leaves α_2 of (2.16) as the leading term in (2.21). With ρ confined to $[-1,1]$, the higher-order moments $\alpha_j = E_g\{\rho^j\}$ often decrease quickly to zero.

The root mean square (rms) correlation

$$\alpha = \alpha_2^{1/2} = \left[\int_{-1}^{1} \rho^2 g(\rho) d\rho \right]^{\frac{1}{2}} \tag{2.22}$$

featured in Efron (2007a) (where it is called the *total correlation*), is a single-number summary of $\boldsymbol{z}_i$'s entire correlation structure. Carrying out the three reductions above produces a greatly simplified form of (2.21),

$$\textit{rms approximation}: \quad \mathbf{cov}_1 \doteq (N\Delta\alpha)^2 \bar{\boldsymbol{\phi}}^{(2)} \bar{\boldsymbol{\phi}}^{(2)'}/2 \tag{2.23}$$

with $\bar{\boldsymbol{\phi}}^{(2)}$ in (2.20) depending on the second derivative of the normal density, $\varphi^{(2)}(u) = \varphi(u)\cdot(u^2-1)$.

Figure 2 compares the exact formulas (2.12)–(2.14) for $\mathbf{cov}(\boldsymbol{y})$ with the simplified formula based on the rms approximation (2.23); for a numerical example having $N = 6000, \alpha = .10$, and two classes (2.4)–(2.5), initially with

$$(\mu_0,\sigma_0) = (0,1),\ p_0 = .95 \quad \text{and} \quad (\mu_1,\sigma_1) = (2.5,1),\ p_1 = .05 \tag{2.24}$$

but then recentered as in the leukemia example; see Remark D of Section 6 for more detail. The plotted curves show the standard deviations $\mathrm{sd}\{y_k\} = \mathrm{cov}_{kk}(\boldsymbol{y})^{\frac{1}{2}}$ from (2.12), the corresponding rms approximation (2.23), and also $\mathrm{sd}_0\{y_k\} = (\mathrm{cov}_{0kk})^{\frac{1}{2}}$ from (2.13). We can see there is a substantial correlation penalty over most of the range of z, and also that the rms approximation is quite satisfactory here.

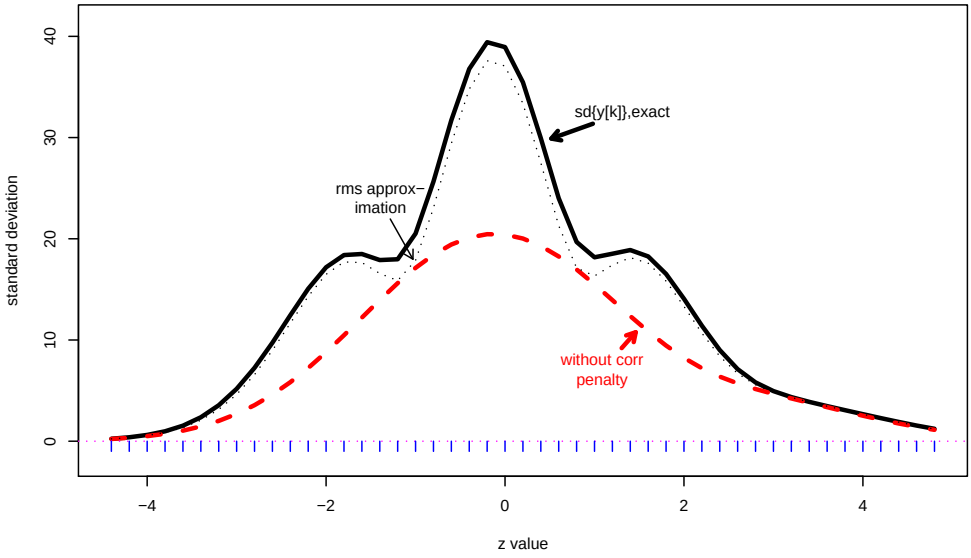

Figure 2: Comparison of exact formula for standard deviation of y_k from (2.12) (heavy curve) with rms approxmation from (2.23) (dotted curve); $N = 6000, \alpha = .10$ in (2.22), two classes as in (2.24). Dashed curve is standard deviation from (2.13) ignoring the correlation penalty. Hash marks indicate bin midpoints x_k.

Returning to right-sided cdfs (2.1), let $\boldsymbol{B}$ be the $K \times K$ matrix

$$\boldsymbol{B}_{kk'} = \begin{cases} 1 & \text{if } k \le k' \\ 0 & \text{if } k > k' \end{cases} \tag{2.25}$$

so

$$\hat{\boldsymbol{F}} = \frac{1}{N}\boldsymbol{B}\boldsymbol{y} \tag{2.26}$$

is a K-vector with kth component the proportion of z_i's in bins indexed $\ge k$,

$$\hat{F}_k = \#\{z_i \ge x_k - \Delta/2\}/N \qquad (k = 1,2,\ldots,K). \tag{2.27}$$

($\boldsymbol{B}$ would be transposed if we were dealing with left-sided cdfs.) The expectation of $\hat{\boldsymbol{F}}$ is both obvious and easy to obtain from (2.9),

$$\begin{aligned} E\left\{\hat{F}_k\right\} &= \sum_c p_c \left[\sum_{k'\geq k} \Delta\varphi\left(\frac{x_{k'}-\mu_c}{\sigma_c}\right)\Big/\sigma_c\right] \doteq \sum_c p_c \int_{x_{kc}}^{\infty} \varphi(u)du \\ &= \sum_c p_c \Phi^+(x_{kc}) \end{aligned} \tag{2.28}$$

where $\Phi^+(u) = 1-\Phi(u)$. Now that we are working with tail areas rather than densities we can let $\Delta \to 0$, making (2.28) exact.

$\hat{\boldsymbol{F}}$ has covariance matrix $\boldsymbol{B}\mathbf{cov}(\boldsymbol{y})\boldsymbol{B}'/N^2$. The same kind of calculations as in (2.28) applied to Lemma 1 gives the following theorem.

Theorem 1. *Under the multiclass model* (2.4)–(2.5)*,*

$$\mathbf{Cov}(\hat{\boldsymbol{F}}) = \mathbf{Cov}_0 + \mathbf{Cov}_1 \tag{2.29}$$

where $\mathbf{Cov}_0$ *has klth entry*

$$\frac{1}{N}\sum_c p_c\left\{\Phi^+\left(\max(x_{kc},x_{lc})\right) - \Phi^+(x_{kc})\Phi^+(x_{lc})\right\} \tag{2.30}$$

and

$$\mathbf{Cov}_1 = \sum_j \frac{\alpha_j}{j!}\bar{\boldsymbol{\varphi}}^{(j-1)}\bar{\boldsymbol{\varphi}}^{(j-1)'} - \frac{1}{N}\sum_j \frac{\alpha_j}{j!}\left\{\sum_c p_c \boldsymbol{\varphi}_c^{(j-1)}\boldsymbol{\varphi}_c^{(j-1)'}\right\}. \tag{2.31}$$

Here p_c is from (2.5), x_{kc} and x_{lc} from (2.6), α_j is as in (2.16) and

$$\bar{\boldsymbol{\varphi}}^{(j-1)} = \sum_c p_c \boldsymbol{\varphi}_c^{(j-1)} = \sum_c p_c \varphi^{(j-1)}(\boldsymbol{x}_c). \tag{2.32}$$

(Notice the distinction between $\bar{\boldsymbol{\varphi}}$ and $\bar{\boldsymbol{\phi}}$ (2.20), and between $\mathbf{Cov}$ and $\mathbf{cov}$ etc., Lemma 1.)

The three-step reduction leading to (2.31) also can be applied to $\mathbf{Cov}_1$: for α as in (2.22),

$$\textit{Rms approximation:}\quad \mathbf{Cov}_1 \doteq \alpha^2\bar{\boldsymbol{\varphi}}^{(1)}\bar{\boldsymbol{\varphi}}^{(1)'}/2 \tag{2.33}$$

with $\bar{\boldsymbol{\varphi}}^{(1)}$ depending on the first derivative of the normal density, $\varphi^{(1)}(u) = -\varphi(u)u$. Section 3 shows that (2.33) is especially convenient for applications.

Figure 3 is the version of Figure 2 applying to $\hat{\boldsymbol{F}}$: the heavy curve tracks $\mathrm{sd}(\hat{F}_k)$ from (2.29), the dotted curve is from Rms approximation (2.33), and the dashed curve shows the standard deviations from $\mathbf{Cov}_0$ (2.30), ignoring the correlation penalty. Once again the simple approximation formula performs well, particularly for extreme values of z, which are likely to be the ones of interest in applications. The correlation penalty is more severe here than in Figure 2, especially in the tails.

The $\mathbf{Cov}_0$ formula (2.30) is essentially the covariance function for a Brownian bridge. Results related to Theorem 1 can be found in the empirical process literature; see equation (2.2) of Csörgő and Mielniczuk (1996) for example, which applies to the "one-class" case when all the z_i's are $\mathcal{N}(0,1)$. Desai et al. (2009) extend the covariance calculations in Efron (2007a) to include skewness corrections.

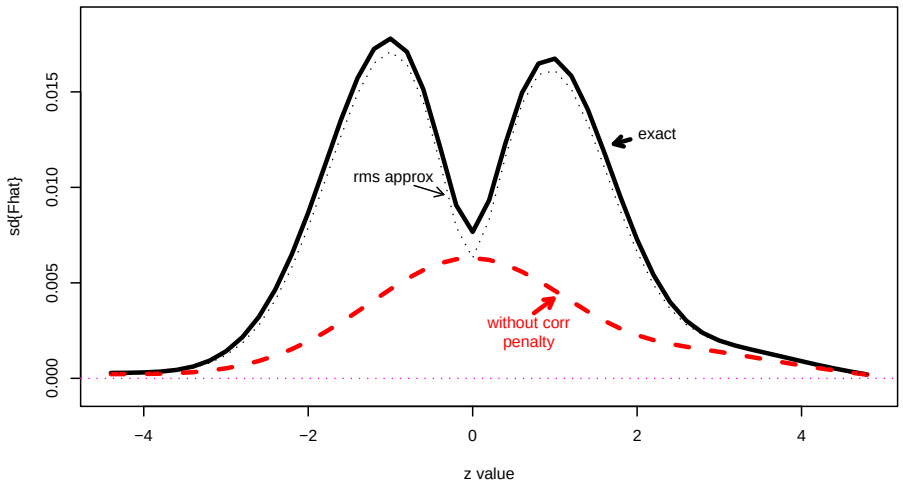

Figure 3: Comparison of exact formula for $\text{sd}\{\hat{F}_k\}$ from Theorem 1 (heavy curve) with Rms approximation using (2.33) (dotted curve); same example as in Figure 2. Dashed curve shows standard deviation estimates ignoring the correlation penalty.

3 Estimation of the correlation parameters

Application of Section 2's theory requires us to estimate several parameters: the rms correlation α (2.22), and the class components (p_c, μ_c, σ_c) in (2.4)–(2.5) (though we will see that the latter task can be avoided under some assumptions). This section illustrates the estimation process in terms of the leukemia study of Section 1. $\boldsymbol{X}$, the data matrix for the study, has $N = 7128$ rows, one for each gene, and $n = 72$ columns, one for each patient; the $n_1 = 47$ ALL patients precede the $n_2 = 25$ AML patients. Entry x_{ij} of $\boldsymbol{X}$ is the expression level for gene i on patient j. The columns of $\boldsymbol{X}$ were individually standardized to have mean 0 and variance 1; see Remark E.

The ith row of $\boldsymbol{X}$ gives t_i, the two-sample t-statistic comparing expression levels on gene i for AML versus ALL patients. These are converted to z-values $z_i = \Phi^{-1}(F_{70}(t_i))$ (1.2), whose histogram appears in Figure 1. As noted before, the histogram is much wider near its center than a theoretical $\mathcal{N}(0,1)$ null distribution: analysis using the `locfdr` program described in Efron (2007b, 2008) estimated that proportion $p_0 = .93$ of the genes were "null" (i.e., identically distributed for ALL and AML), and that z-values for the null genes followed a $\mathcal{N}(.09, 1.68^2)$ distribution.

We wish to estimate the rms correlation α (2.22). Let $\boldsymbol{X}_0$ indicate an $N \times n_0$ subset of $\boldsymbol{X}$ pertaining to a single population of subjects, for example the 47 ALL patients. There are $N \cdot (N-1)/2$ sample correlations $\hat{\rho}_{ii'}$ between rows i and i' of $\boldsymbol{X}_0$. Computing all of these, or a sufficiently large random sample, yields the empirical mean and variance (m, v) of the $\hat{\rho}$ distribution,

$$\hat{\rho} \sim (m, v), \tag{3.1}$$

$(m, v) = (.002, .190^2)$ for the ALL patients. As discussed in Section 3 of Efron (2009), standardizing the columns of $\boldsymbol{X}_0$ to have mean 0 forces $m \doteq 0$, and we will assume $m = 0$

in what follows. (This is equivalent to taking $\alpha_1 = 0$ as we did following (2.21).)

The obvious choice $\bar{\alpha} = v^{\frac{1}{2}}$ tends to greatly overestimate α: each $\hat{\rho}_{ii'}$ is nearly unbiased for its true correlation $\rho_{ii'}$, a normal-theory approximation for mean and variance being

$$\hat{\rho}_{ii'} \sim \big(\rho_{ii'}, (1-\rho_{ii'}^2)^2/(n-3)\big) \tag{3.2}$$

(Johnson and Kotz, 1970), but the considerable variances in (3.2) can greatly broaden the empirical distribution of the $\hat{\rho}$'s. Two corrected estimates of α are developed in Efron (2009). The simpler correction formula is

$$\hat{\alpha}^2 = \frac{n_0}{n_0 - 1}\left(v - \frac{1}{n_0 - 1}\right) \tag{3.3}$$

based on an identity between the row and column correlations of $\boldsymbol{X}_0$. The second approach uses an empirical Bayes analysis of the variance term in (3.3) to justify a more elaborate formula,

$$\tilde{\alpha}^2 = \tilde{v} - \frac{3}{n-5}\tilde{v}^2 \qquad \left[\tilde{v} = \frac{(n-3)v - 1}{n-5}\right]. \tag{3.4}$$

The first three columns of Table 2 compare $\hat{\alpha}$ with $\tilde{\alpha}$ for $\boldsymbol{X}_0$ based on the ALL patients, the AML patients, and both. The final column reports mean $\pm$ standard deviation for $\hat{\alpha}$ and $\tilde{\alpha}$ in 100 simulations of model (2.24): $N = 6000, n_1 = n_2 = 40$ patients in each class, true $\alpha = .10$; see Remark D. The two estimates are effectively linear functions of each other for typical values of v; $\hat{\alpha}$, the simpler choice, is preferred by the author.

	ALL	AML	Both	Simulation
$\hat{\boldsymbol{\alpha}}$:	.121	.109	.114	$.1054 \pm .0074$
$\tilde{\boldsymbol{\alpha}}$:	.118	.092	.113	$.1045 \pm .0075$

Table 2: Estimates $\hat{\alpha}$ and $\tilde{\alpha}$, (3.3) and (3.4), for rms correlation α (2.22) of leukemia data; also 100 simulations of model (2.24), $N = 6000, n_1 = n_2 = 40$, true $\alpha = .10$, showing mean $\pm$ standard deviation.

It seems that we need to estimate the class components (p_c, μ_c, σ_c) in (2.4)–(2.5) in order to apply the theory of Section 2, but under certain assumptions this can be finessed, as discussed next.

The marginal density $f(z)$ under model (2.4)–(2.5) is

$$f(z) = \sum_c p_c \varphi\left(\frac{z - \mu_c}{\sigma_c}\right)\frac{1}{\sigma_c}; \tag{3.5}$$

so, letting $\boldsymbol{f} = f(\boldsymbol{x})$ (the density evaluated at the K-vector of bin midpoints), we have $\Delta \cdot \boldsymbol{f} = \sum_c p_c \boldsymbol{\pi}_c$ as in (2.8). Formula (2.13) can be expressed as

$$\mathbf{cov}_0 = N\left\{\mathrm{diag}(\Delta \boldsymbol{f}) - \sum_c p_c \boldsymbol{\pi}_c \boldsymbol{\pi}_c'\right\}. \tag{3.6}$$

Here we are assuming, as in (2.5), that the class sample sizes N_c are fixed. A more realistic assumption might be that the numbers $N_1, N_2, \ldots, N_C$ are a multinomial sample of size N, sampled with probabilities $p_1, p_2, \ldots, p_C$, in which case (3.6) becomes the usual multinomial covariance matrix

$$\mathbf{cov}_0 = N\left\{\text{diag}(\Delta \boldsymbol{f}) - \Delta^2 \boldsymbol{f}\boldsymbol{f}'\right\}. \tag{3.7}$$

A smooth curve $\hat{\boldsymbol{f}}$ fit to the histogram heights[2] as in Figure 1 then yields $\widehat{\mathbf{cov}}_0$ by substitution into (3.7), *without requiring knowledge of the class structure* (2.4)–(2.5). In the same way, we can estimate the $\mathbf{Cov}_0$ for $\hat{\boldsymbol{F}}$ in (2.30) by the standard multinomial formula

$$\left(\widehat{\mathbf{Cov}}_0\right)_{kl} = \frac{1}{N}\left\{\hat{F}_{\max(k,l)} - \hat{F}_k\hat{F}_l\right\}. \tag{3.8}$$

Under some circumstances, a similar tactic can be applied to estimate the correlation penalties $\mathbf{cov}_1$ and $\mathbf{Cov}_1$, (2.23) and (2.33). The first and second derivatives $f^{(1)}(z)$ and $f^{(2)}(z)$ of (3.5) are

$$f^{(1)}(z) = \sum_c p_c\varphi^{(1)}\left(\frac{z-\mu_c}{\sigma_c}\right)\frac{1}{\sigma_c^2} \quad \text{and} \quad f^{(2)}(z) = \sum_c p_c\varphi^{(2)}\left(\frac{z-\mu_c}{\sigma_c}\right)\frac{1}{\sigma_c^3}. \tag{3.9}$$

Suppose we make the *homogeneity assumption* that all σ_c values are the same, say $\sigma_c = \sigma_0$. Comparison with definitions (2.32) and (2.20) then gives

$$\bar{\boldsymbol{\varphi}}^{(1)} = \sigma_0^2\boldsymbol{f}^{(1)} \quad \text{and} \quad \bar{\boldsymbol{\phi}}^{(2)} = \sigma_0^2\boldsymbol{f}^{(2)} \tag{3.10}$$

with $\boldsymbol{f}^{(j)} = (f^{(j)}(x_k))'$. This leads to the convenient covariance penalty formulas,

$$\mathbf{Cov}_1 \doteq \frac{(\sigma_0^2\alpha)^2}{2}\boldsymbol{f}^{(1)}\boldsymbol{f}^{(1)\prime} \quad \text{and} \quad \mathbf{cov}_1 \doteq \frac{(N\Delta\sigma_0^2\alpha)^2}{2}\boldsymbol{f}^{(2)}\boldsymbol{f}^{(2)\prime} \tag{3.11}$$

from (2.33) and (2.23).

A smooth estimate $\hat{f}(z)$ of $f(z)$ can be differentiated to give estimated values of $\mathbf{Cov}_1$ and $\mathbf{cov}_1$, for example,

$$\text{sd}_1\left\{\hat{F}_k\right\} = \left(\widehat{\mathbf{Cov}}_1\right)_{kk}^{\frac{1}{2}} = \frac{\hat{\sigma}_0^2\hat{\alpha}}{\sqrt{2}}\left|\hat{f}^{(1)}(x_k)\right| \tag{3.12}$$

for the correlation penalty standard deviation of $\hat{F}(x_k)$ (2.27). (This provides the second term in formula (1.4).) The heavy curve in Figure 4 shows (3.12) for the leukemia data, using $\hat{\sigma}_0 = 1.68, \hat{\alpha} = .114$, and $\hat{f}(z)$ from Figure 1.

Suppose we are unwilling to make the homogeneity assumption. A straightforward approach to estimating $\mathbf{Cov}_1$ or $\mathbf{cov}_1$ requires assessments of the parameters (p_c, μ_c, σ_c) in (2.4)–(2.5). These can be based on the "non-null counts" (Efron, 2007b), the small bars plotted negatively in Figure 1; see Remark B. The figure suggests three classes, left, center and right, with parameter values as estimated in Table 3.

[2]The estimate used here is a Poisson spline regression as described following (4.10).

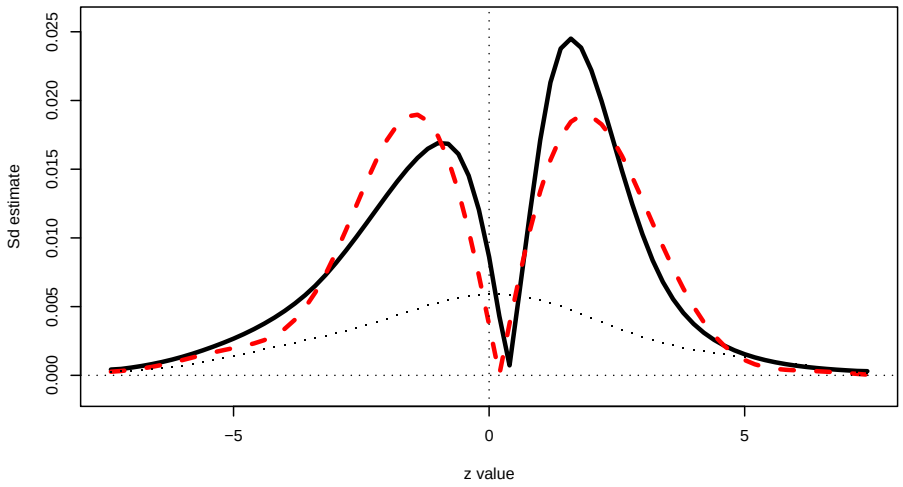

Figure 4: Leukemia data; two estimates of correlation penalty standard deviation $\mathrm{sd}_1\{\hat{F}_k\}$ for $\hat{F}_k$ (2.27). *Solid curve* formula (3.12); *dashed curve* Rms approximation (2.33) using class estimates from Table 3. *Dotted curve* is independence estimate from (3.8), indicating that the correlation penalty is substantial.

	left	center	right
$\boldsymbol{p}_c$:	.054	.930	.016
$\boldsymbol{\mu}_c$:	−4.2	.09	5.4
$\boldsymbol{\sigma}_c$:	1.16	1.68	1.05

Table 3: Three-class model (2.4)–(2.5) for leukemia data. Parameter estimates based on non-null counts, Remark B.

The dashed curve in Figure 4 shows $\mathrm{sd}_1(\hat{F}_k)$ estimated directly from (2.32)–(2.33) using the values in Table 3. It is similar to the homogeneity estimate (3.12) except in the extreme tails.

Formula (1.4) for the standard deviation of $\hat{F}(x)$ was tested in a simulation experiment. The specifications were the same as in Figure 3, with $N = 6000$, $\alpha = .10$, and two classes of z-values (2.24). One hundred $\boldsymbol{X}$ matrices were generated as in the simulation for Table 2, each yielding a vector of 6000 correlated z-values, followed by $\hat{\sigma}_0$, $\hat{\alpha}$, and $\hat{f}^{(1)}(x)$ for use in (1.4); see Remark D for further details. Finally, $\widehat{\mathrm{sd}}$, the square root of (1.4), was calculated along with $\widehat{\mathrm{sd}}_0$, the square root of just the first term.

The solid curve in Figure 5 shows the average of the $\widehat{\mathrm{sd}}$ values for x between -4 and 4.5, with solid bars indicating standard deviations of the 100 $\widehat{\mathrm{sd}}$'s. There is a good match of the average with the exact sd curve from Figure 3. The error bars indicate moderate variability across the replications. The average for $\widehat{\mathrm{sd}}_0$, dashed curve, agrees with the corresponding curve in Figure 3 and shows that correlation cannot be ignored in this situation.

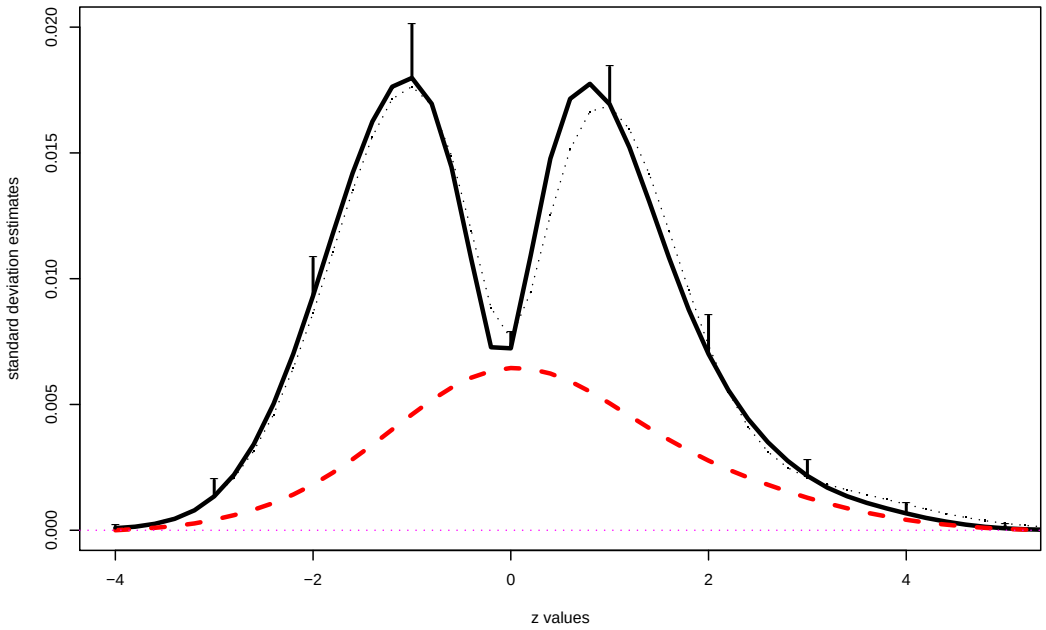

Figure 5: Simulation experiment for formula (1.4). *Solid curve* average of $\widehat{\text{sd}}$, square root of (1.4), 100 replications, with bars indicating standard deviation of $\widehat{\text{sd}}$ at $x = -4, -3, \ldots, 4$; *dotted curve* exact sd from Figure 3; *dashed curve* average of $\widehat{\text{sd}}_0$, standard error estimate for $\hat{F}(x)$ ignoring correlation.

4 Applications

Correlation usually degrades statistical accuracy, an important question for the data analyst being the severity of its effects on the estimates and tests at hand. The purpose of Section 2 and Section 3 was to develop practical methods for honestly assessing the accuracy of inferences made in the presence of large-scale correlation. This section presents a few examples of the methodology in action.

We have already seen one example: in Table 1 the accuracy of $\hat{F}(x)$, the right-sided empirical cdf for the leukemia data, computed from the usual binomial formula that assumes independence among the z-values,

$$\widehat{\text{sd}}_0 = \left\{\hat{F}(x)\left(1-\hat{F}(x)\right)\Big/N\right\}^{\frac{1}{2}}, \tag{4.1}$$

was compared with $\widehat{\text{sd}}$ from formula (1.4) in which the correlation penalty term was included: $\widehat{\text{sd}}$ more than doubled $\widehat{\text{sd}}_0$ over most of the range.

Suppose we assume, as in Efron (2008), that each of the N cases (the N genes in the leukemia study) is either *null* or *non-null* with prior probability p_0 or $p_1 = 1 - p_0$, and with the corresponding z-values having density either $f_0(z)$ or $f_1(z)$,

$$\begin{aligned} p_0 &= \Pr\{\text{null}\} \qquad & f_0(z) \text{ density if null,} \\ p_1 &= \Pr\{\text{non-null}\} \qquad & f_1(z) \text{ density if non-null.} \end{aligned} \tag{4.2}$$

Let F_0 and F_1 be the right-sided cdfs of f_0 and f_1, and F the mixture cdf

$$F(x) = p_0 F_0(x) + p_1 F_1(x). \tag{4.3}$$

The probability of a case being null given that z exceeds x is

$$\text{Fdr}(x) \equiv \Pr\{\text{null}|z \geq x\} = \frac{p_0 F_0(x)}{F(x)} \tag{4.4}$$

according to Bayes theorem, "Fdr" standing for false discovery rate.

If p_0 and F_0 are known then Fdr has the obvious estimate

$$\widehat{\text{Fdr}}(x) = p_0 F_0(x)/\hat{F}(x) \tag{4.5}$$

(2.1). Benjamini and Hochberg's celebrated 1995 algorithm uses $\widehat{\text{Fdr}}(x)$ for simultaneous hypothesis testing, but it can also be thought of as an empirical Bayes estimator of the Bayesian probability $\text{Fdr}(x)$. The bottom row of Table 1 shows $\widehat{\text{Fdr}}(x)$ for the leukemia data, taking $p_0 = .93$ and $F_0 \sim \mathcal{N}(.09, 1.68^2)$ as in Figure 1. (Later we will do a more ambitious calculation taking into account the estimation of p_0 and F_0.)

The coefficient of variation for $\widehat{\text{Fdr}}(x)$ approximately equals that for $\hat{F}(x)$ (when $p_0 F_0(x)$ is known in (4.5)). At $x = 5$ we have $\widehat{\text{Fdr}}(5) = .15$, with coefficient of variation about .19. An $\widehat{\text{Fdr}}$ of .15 might be considered small enough to trigger significance in the Benjamini–Hochberg algorithm, but in any case it seems clear that the probabiltiy of being null is quite low for the 71 genes having z_i above 5. Even taking account of correlation effects, we have a rough upper confidence limit of .21 (i.e., $.15 \cdot (1 + 2 \cdot .19)$) for $\text{Fdr}(5)$.

Next we consider accuracy estimates for a general class of statistics $Q(\boldsymbol{y})$, where Q is a q-dimensional function of the count vector $\boldsymbol{y}$ (2.3). As in Section 5 of Efron (2007b), we assume that a small change $d\boldsymbol{y}$ in the count vector (considered as varying continuously) produces change dQ in Q according to

$$dQ = \hat{\boldsymbol{D}} d\boldsymbol{y} \qquad \left[\hat{D}_{jk} = \partial Q_j/\partial y_k\right]. \tag{4.6}$$

If $\widehat{\mathbf{cov}}(\boldsymbol{y})$ is a covariance estimate for $\boldsymbol{y}$, obtained perhaps as in (2.12), (3.8), or (3.11), then the usual delta-method estimate for $\mathbf{cov}(Q)$ is

$$\widehat{\mathbf{cov}}(Q) = \hat{\boldsymbol{D}} \widehat{\mathbf{cov}}(\boldsymbol{y}) \hat{\boldsymbol{D}}'. \tag{4.7}$$

In a theoretical context, where $\mathbf{cov}(\boldsymbol{y})$ is known, we might instead use

$$\text{cov}(Q) \doteq \boldsymbol{D}\mathbf{cov}(\boldsymbol{y})\boldsymbol{D}' \tag{4.8}$$

now with the derivative matrix $\boldsymbol{D}$ evaluated at the expectation of $\boldsymbol{y}$.

Model (4.2) yields the *local false discovery rate*

$$\text{fdr}(x) \equiv \Pr\{\text{null}|z = x\} = p_0 f_0(x)/f(x), \tag{4.9}$$

$f(x)$ being the mixture density

$$f(x) = p_0 f_0(x) + p_1 f_1(x); \tag{4.10}$$

$\text{fdr}(x)$ is inferentially more appropriate than $\text{Fdr}(x)$ from a Bayesian point of view, but it is not as immediately available since it involves estimating the *density* $f(x)$. However, because z-value densities are mixtures of near-normals as shown in Section 5, it is usually straightforward to carry out the estimation.

`Locfdr`, the algorithm discussed in Efron (2007a, 2008), estimates $f(x)$ by means of Poisson regression of the counts y_k as a spline function of the x_k, the bin midpoints in (2.2)–(2.3). The structure matrix $\boldsymbol{M}$ for the Poisson regression is $K \times d$, where K is the number of bins and d is degrees of freedom (e.g., the number of free parameters of the spline fit; see Remark G for details). Let $\hat{\boldsymbol{f}}$ be the vector of fitted values $\hat{f}(x_k)$, and $\hat{\boldsymbol{\ell}}$ the vector with components $\hat{\ell}_k = \log(\hat{f}(x_k))$. Then, as discussed in (Efron, 2007b, Sect. 5), (4.6) takes the form

$$d\hat{\boldsymbol{\ell}} = \hat{\boldsymbol{D}} d\boldsymbol{y} \qquad \text{with } \hat{\boldsymbol{D}} = \boldsymbol{M}\left(\boldsymbol{M}' \operatorname{diag}\left(N\Delta\hat{\boldsymbol{f}}\right)\boldsymbol{M}\right)^{-1}\boldsymbol{M}' \tag{4.11}$$

and we can use (4.7) or (4.8) to approximate $\mathbf{cov}(\hat{\boldsymbol{\ell}})$.

For any function $v(x)$ define the vector

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_k, \ldots, v_K)' = (\ldots, v(x_k), \ldots)' \tag{4.12}$$

as with $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{\ell}}$ above. If

$$\widehat{\text{lfdr}}(x) \equiv \log\left(\widehat{\text{fdr}}(x)\right) = \log\left(p_0 f_0(x)\right) - \log\left(f(x)\right) \tag{4.13}$$

then

$$\widehat{\mathbf{lfdr}}(x) = \log(p_0) + \log(\boldsymbol{f}_0) - \hat{\boldsymbol{\ell}} \tag{4.14}$$

implying, if p_0 and f_0 are known, that

$$\mathbf{cov}\left(\widehat{\mathbf{lfdr}}(x)\right) = \operatorname{cov}(\hat{\boldsymbol{\ell}}) \doteq \boldsymbol{D}\mathbf{cov}(\boldsymbol{y})\boldsymbol{D}' \tag{4.15}$$

with $\boldsymbol{D} = \boldsymbol{M}(\boldsymbol{M}' \operatorname{diag}(N\Delta\boldsymbol{f})\boldsymbol{M})^{-1}\boldsymbol{M}'$ (4.8).

The solid curves in Figure 6 plot standard deviations for $\log(\widehat{\text{fdr}}(x))$, obtained as square root of the diagonal elements of $\mathbf{cov}(\widehat{\mathbf{lfdr}})$ (4.15), for model (2.24) with $N = 6000$ and rms correlation α equal $0, .1$, or $.2$; see Remark D. The horizontal axes are plotted in terms of the upper percentiles of $F(x)$, the right end of each plot corresponding to the far right tail of the z-value distribution. For $\alpha = 0$, $\text{sd}(\log\widehat{\text{fdr}}(x))$ increases from .03 to .08 as we move from the fifth to the first percentile of F. The coefficient of variation (CV) of $\widehat{\text{fdr}}(x)$ nearly equals $\text{sd}(\log\widehat{\text{fdr}}(x))$, so $\widehat{\text{fdr}}(x)$ is quite accurately estimated for $\alpha = 0$, but substantially less so for $\alpha = .2$. Reducing N to 1500 doubles the standard deviation estimates for $\alpha = 0$, but has less effect in the correlated situations: for $\alpha = .1$ for example, the increase is only 20% at percentile .025. Simulations confirmed the correctness of these results.

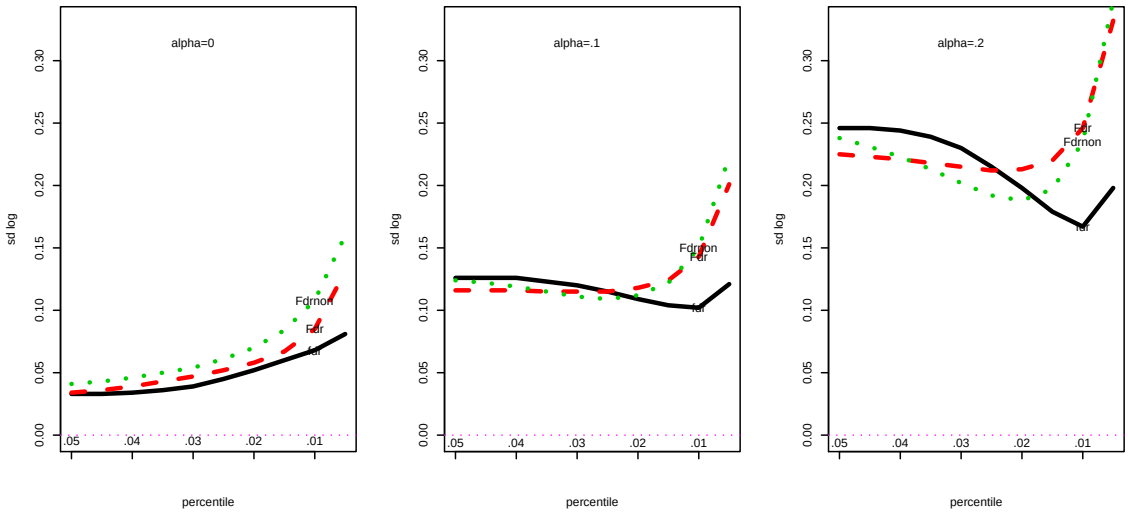

Figure 6: *Solid curves* show standard deviation of $\log(\widehat{\text{fdr}}(x))$ as a function of x at the upper percentiles of the z-value distribution for model (2.24), $N = 6000$ and $\alpha = 0, .1, .2$. *Dotted curves* (green) same for $\log(\widehat{\text{Fdr}}(x))$ (4.5), nonparametric Fdr estimator. *Dashed curves* (red) for parametric version (4.17) of Fdr estimator.

Intuitively it seems that fdr should be harder to estimate than Fdr, but that is not what Figure 6 shows. Let $\hat{L}_k = \log(\hat{F}(x_k))$, with corresponding vector $\hat{\boldsymbol{L}}$. Then $\hat{\boldsymbol{D}}$ in (4.6) has

$$\hat{D}_{jk} = B_{jk} / \left(N \cdot \hat{F}_j \right) \tag{4.16}$$

with $\boldsymbol{B}$ as in (2.25), giving an estimate of $\mathbf{cov}(\hat{\boldsymbol{L}})$ from (4.7) or (4.8). The same argument as (4.13)–(4.15) shows that this also estimates $\mathbf{cov}(\log \widehat{\mathbf{Fdr}})$, the log of vector (4.5), assuming $p_0 F_0(x)$ is known. The dotted curves in Figure 6 show standard deviations for $\log(\widehat{\text{Fdr}}(x))$. If anything, Figure 6 suggests that $\widehat{\text{fdr}}$ is *less* variable than $\widehat{\text{Fdr}}$, particularly at the smaller percentiles.

Here we are comparing the nonparametric estimator $\widehat{\text{Fdr}}(x)$ (4.5) with the parametric estimator $\widehat{\text{fdr}}(x)$. The Poisson spline estimate $\hat{f}(x)$ that gave $\widehat{\text{fdr}}(x)$ can be summed to give parametric estimates of $F(x)$ and $\text{Fdr}(x)$, say $\widetilde{\text{Fdr}}(x)$. Straightforward calculations show that the derivative matrix $\hat{\boldsymbol{D}}$ for $\widetilde{\text{Fdr}}(x)$ is

$$\hat{\boldsymbol{D}} = \hat{\boldsymbol{C}} \hat{\boldsymbol{D}}_f \qquad \text{where } C_{jk} = B_{jk} \hat{f}_k / \hat{F}_j \tag{4.17}$$

with $\boldsymbol{B}$ from (2.25) and $\hat{\boldsymbol{D}}_f$ equaling $\hat{\boldsymbol{D}}$ in (4.11). Standard deviations for $\log(\widetilde{\text{Fdr}})$, shown by the dashed curves in Figure 6, indicate about the same accuracy for $\widetilde{\text{Fdr}}(x)$ as for $\widehat{\text{Fdr}}$.

All of these calculations assumed that p_0 and $f_0(z)$ (or $F_0(z)$) in (4.2) were known. This is unrealistic in situations like the leukemia study, where there is clear evidence that a textbook $\mathcal{N}(0, 1)$ theoretical null distribution is too narrow. Estimating an "empirical null" distribution, such as $\mathcal{N}(.09, 1.68^2)$ in Figure 1, is both necessary and feasible (see Efron, 2008) but can greatly increase variability, as discussed next.

Formula (4.14) becomes

$$\widehat{\mathbf{lfdr}} = \log(\hat{p}_0) + \log\left(\hat{\boldsymbol{f}}_0\right) - \hat{\boldsymbol{\ell}} \tag{4.18}$$

when p_0 and f_0 are themselves estimated. The corresponding derivative matrix $\hat{\boldsymbol{D}} = d\widehat{\mathbf{lfdr}}/d\boldsymbol{y}$ in (4.6) appears as equation (5.8) in Efron (2007b), this formula applying to the *central matching* method for estimating $p_0 f_0(z)$. The second row of Table 4 shows $\mathrm{sd}\{\log\widehat{\mathrm{fdr}}(x)\}$ obtained from $\boldsymbol{D}\mathrm{cov}(\boldsymbol{y})\boldsymbol{D}'$ for the same situation as in the middle panel of Figure 6. Comparison with the theoretical null standard deviations (from the solid curve in the middle panel) shows that estimating the null distribution greatly increases variability.

percentile:	0.05	0.04	0.03	0.02	0.01
sd empirical null:	**0.18**	**0.26**	**0.36**	**0.54**	**0.83**
sd theoretical null:	0.13	0.13	0.12	0.11	0.10
x:	1.98	2.16	2.40	2.74	3.25
$\mathrm{fdr}(x)$:	0.69	0.58	0.44	0.25	0.09
$\mathrm{Fdr}(x)$:	0.34	0.27	0.19	0.10	0.04

Table 4: Comparison of $\mathrm{sd}\{\log(\widehat{\mathrm{fdr}}(x))\}$ using empirical null versus theoretical null for the situation in the middle panel of Figure 6. The empirical null standard deviations are much larger, as seen also in Efron (2007b).

Here are some points to note:

- Accuracy is worse for $\log(\widehat{\mathrm{Fdr}})$ then for $\log(\widehat{\mathrm{fdr}})$ in the top line of Table 4.
- Accuracy is somewhat better when $p_0 f_0(z)$ is estimated by the MLE option in `locfdr` (Lemma 2 of Efron, 2007b).
- The big empirical null standard deviations in Table 4 are at least partially misleading: some of the variability in $\widehat{\mathrm{fdr}}(x)$ is "signal" rather than "noise", tracking conditional changes in the appropriate value of $\mathrm{fdr}(x)$. See Figure 2 of Efron (2007a) and the discussion in that paper.

Remark H of Section 6 describes a parametric bootstrap resampling scheme that avoids the Taylor series computations of (4.7), but which has not yet been carefully investigated.

5 The non-null distribution of z-values

The results of the previous sections depend on the variates z_i having normal distributions (1.5). By definition, a z-value is a statistic having a $\mathcal{N}(0,1)$ distribution under a null hypothesis H_0 of interest (1.1): but will it still be normal for non-null conditions? This

section shows that under repeated sampling the non-null distribution of z will typically have mean $O(1)$, standard deviation $1+O(n^{-\frac{1}{2}})$, and non-normality $O_p(n^{-1})$ (as measured by the magnitude of skewness and kurtosis). In other words, normality degrades more slowly than unit standard deviation as we move away from the null hypothesis.

Figure 7 illustrates the phenomenon for the case of non-central t distributions,

$$z = \Phi^{-1}\left(F_\nu(t)\right) \qquad t \sim t_\nu(\delta), \tag{5.1}$$

the notation indicating a non-central t variable with ν degrees of freedom and non-centrality parameter δ (*not* δ^2), as described in Chapter 31 of Johnson and Kotz (1970). Here, as in (1.2), F_ν is the cdf of a *central* t_ν distribution. The standard deviation of z decreases as $|\delta|$ increases; for $\delta = 5, \nu = 20$, z has (mean, sd) equal $(4.01, 0.71)$. The useful and perhaps surprising observation is that normality holds up quite well even far from the null case $\delta = 0$. We tacitly used this fact to justify application of our theoretical results to the leukemia study.

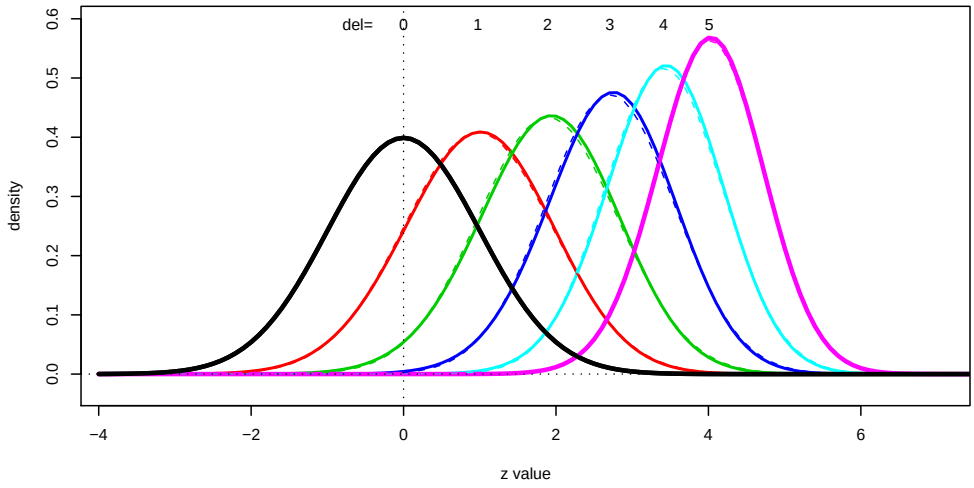

Figure 7: Density of the z-value statistic (5.1) when t has a noncentral t distribution with $\nu = 20$ degrees of freedom; for non-centrality parameter $\delta = 0, 1, 2, 3, 4, 5$. The densities are seen to be nearly normal; dashed curves are exact normal densities matched in mean and standard deviation. For $\delta = 5$, z has (mean,sd, skew,kurt) $= (4.01, .71, -.06, .08)$. Negative values of δ give mirror image results. Remark G of Section 6 describes the density function calculations.

To begin the theoretical development, suppose that $y_1, y_2, \ldots, y_n$ are independent and identically distributed (iid) observations sampled from F_θ, a member of a one-parameter family of distributions,

$$\mathcal{F} = \{F_\theta, \theta \in \Theta\} \tag{5.2}$$

having its moment parameters {mean, standard deviation, skewness, kurtosis}, denoted

$$\{\mu_\theta, \sigma_\theta, \gamma_\theta, \delta_\theta\}, \tag{5.3}$$

defined differentiably in θ. The results that follow are heuristic in the sense that they only demonstrate second-order Cornish–Fisher expansion properties, with no attempt to provide strict error bounds.

Under the null hypothesis $H_0 : \theta = 0$, which we can write as

$$H_0 : y \sim \{\mu_0, \sigma_0, \gamma_0, \delta_0\}, \tag{5.4}$$

the standardized variate

$$Y_0 = \sqrt{n}\left(\frac{\bar{y} - \mu_0}{\sigma_0}\right) \qquad \left[\bar{y} = \sum_{i=1}^{n} y_i/n\right] \tag{5.5}$$

satisfies

$$H_0 : Y_0 \sim \left\{0, 1, \frac{\gamma_0}{\sqrt{n}}, \frac{\delta_0}{n}\right\}. \tag{5.6}$$

Normality can be improved to second order by means of a Cornish–Fisher transformation,

$$Z_0 = Y_0 - \frac{\gamma_0}{6\sqrt{n}}\left(Y_0^2 - 1\right) \tag{5.7}$$

which reduces the skewness in (5.6) from $O(n^{-\frac{1}{2}})$ to $O(n^{-1})$,

$$H_0 : Z_0 \sim \{0, 1, 0, 0\} + O\left(n^{-1}\right). \tag{5.8}$$

See Chapter 1 of Johnson and Kotz (1970) or, for much greater detail, Section 2.2 of Hall (1992). We can interpret (5.8) as saying that Z_0 is a *second-order z-value*,

$$H_0 : Z_0 \sim \mathcal{N}(0, 1) + O_p\left(n^{-1}\right), \tag{5.9}$$

e.g., *a test statistic giving standard normal p-values accurate to* $O(n^{-1})$.

Suppose now that H_0 is false, and instead H_1 is true, with $y_1, y_2, \ldots, y_n$ iid according to

$$H_1 : y \sim \{\mu_1, \sigma_1, \gamma_1, \delta_1\} \tag{5.10}$$

rather than (5.4). Setting

$$Y_1 = \sqrt{n}\left(\frac{\bar{y} - \mu_1}{\sigma_1}\right) \quad \text{and} \quad Z_1 = Y_1 - \frac{\gamma_1}{6\sqrt{n}}\left(Y_1^2 - 1\right) \tag{5.11}$$

makes Z_1 second-order normal under H_1,

$$H_1 : Z_1 \sim \mathcal{N}(0, 1) + O_p\left(n^{-1}\right). \tag{5.12}$$

We wish to calculate the distribution of Z_0 (5.7) under H_1. Define

$$c = \sigma_1/\sigma_0, \quad d = \sqrt{n}(\mu_1 - \mu_0)/\sigma_0, \quad \text{and} \quad g_0 = \gamma_0/\left(6\sqrt{n}\right). \tag{5.13}$$

Some simple algebra yields the following relationship between Z_0 and Z_1.

Lemma 3. *Under definitions* (5.7), (5.11) *and* (5.13),

$$Z_0 = M + SZ_1 + g_0\left\{\left(\frac{\gamma_1}{\gamma_0}S - c^2\right)\left(Y_1^2 - 1\right) + (1 - c^2)\right\} \tag{5.14}$$

where

$$M = d\cdot(1 - dg_0) \quad \textit{and} \quad S = c\cdot(1 - 2dg_0)\,. \tag{5.15}$$

The asymptotic relationships claimed at the start of this section are easily derived from Lemma 3. We consider a sequence of alternatives θ_n approaching the null hypothesis value θ_0 at rate $n^{-\frac{1}{2}}$,

$$\theta_n - \theta_0 = O\left(n^{-\frac{1}{2}}\right). \tag{5.16}$$

The parameter $d = \sqrt{n}(\mu_{\theta_n} - \mu_0)/\sigma_0$ defined in (5.13) is then of order $O(1)$, as is

$$M = d(1 - dg_0) = d\left(1 - d\gamma_0/\left(6\sqrt{n}\right)\right), \tag{5.17}$$

while standard Taylor series calculations give

$$c = 1 + \frac{\dot\sigma_0}{\dot\mu_0}\frac{d}{\sqrt{n}} + O\left(n^{-1}\right) \quad \text{and} \quad S = 1 + \left(\frac{\dot\sigma_0}{\dot\mu_0} - \frac{\gamma_0}{3}\right)\frac{d}{\sqrt{n}} + O\left(n^{-1}\right), \tag{5.18}$$

the dot indicating differentiation with respect to θ.

Theorem 2. *Under model* (5.2), (5.16), *and the assumptions of Lemma 3,*

$$Z_0 \sim \mathcal{N}(M, S^2) + O_p\left(n^{-1}\right) \tag{5.19}$$

with M and S as given in (5.17)–(5.18). *Moreover,*

$$\left.\frac{dS}{dM}\right|_{\theta_0} = \frac{1}{\sqrt{n}}\left(\left.\frac{d\sigma}{d\mu}\right|_{\theta_0} - \frac{\gamma_0}{3}\right) + O\left(n^{-1}\right). \tag{5.20}$$

Proof. The proof of Theorem 2 uses Lemma 3, with θ_n playing the role of H_1 in (5.14). Both $1 - c^2$ and $(\gamma_1/\gamma_0)S - c^2$ are of order $O(n^{-\frac{1}{2}})$; the former from (5.18) and the latter using $\gamma_1/\gamma_0 = 1 + (\dot\gamma_0/\gamma_0)(\theta_n - \theta_0) + O(n^{-1})$. Since $Y_1^2 - 1$ is $O_p(1)$, this makes the bracketed term in (5.14) $O_p(n^{-\frac{1}{2}})$; multiplying by $g_0 = \gamma_0/(6\sqrt{n})$ reduces it to $O_p(n^{-1})$, and (5.19) follows from (5.12). Differentiating M and S in (5.17)–(5.18) with respect to d verifies (5.20). ■

Theorem 2 supports our claim that, under non-null alternatives, the null hypothesis normality of Z_0 degrades more slowly than its unit standard deviation, the comparison being $O_p(n^{-1})$ versus $O(n^{-\frac{1}{2}})$.

One-parameter exponential families are an important special case of (5.2). With θ the natural parameter of $\mathcal{F}$ and y its sufficient statistic, i.e., with densities proportional to $\exp\{\theta y\}g_0(y)$, (5.20) reduces to

$$\left.\frac{dS}{dM}\right|_{\theta_0} = \frac{\gamma_0}{6\sqrt{n}} + O\left(n^{-1}\right). \tag{5.21}$$

The parameter $\gamma_0/(6\sqrt{n})$ is called the *acceleration* in Efron (1987), interpreted as "the rate of change of standard deviation with respect to expectation on the normalized scale," which agrees with its role in (5.21).

As an example, suppose $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \theta\Gamma_1$, Γ_n indicating a standard gamma distribution with n degrees of freedom, so $g_\theta(y) = (1/\theta)\exp(y/\theta)$ for $y \geq 0$. (Equivalently, $\bar{y} \sim \theta\Gamma_n/n$.) This is an exponential family having skewness $\gamma_0 = 2$ for any choice of θ_0. An exact z-value for testing $H_0 : \theta = \theta_0$ is

$$Z_0 = \Phi^{-1}\left(G_n\left(n\bar{y}/\theta_0\right)\right) \tag{5.22}$$

where G_n is the cdf of Γ_n. Table 5 shows the mean, standard deviation, skewness and kurtosis of Z_0 for $n = 10, \theta_0 = 1$, evaluated for several choices of the alternative θ_1. The standard deviation of Z_0 increases steadily with θ_1; here $\gamma_0/(6\sqrt{n}) = .1054$, matching to better than three decimal places the observed numerical derivative dS/dM. Skewness and kurtosis are both very small; in the equivalent of Figure 7, there is no visible discrepancy at all between the density curves for Z_0 and their matching normal equivalents.

θ_1:	0.4	0.5	0.67	1	1.5	2.0	2.5
mean	−2.49	−1.94	−1.19	0	1.36	2.45	3.38
stdev	0.76	0.81	0.88	1	1.15	1.27	1.38
skew	−0.05	−0.04	−0.02	0	0.02	0.04	0.04
kurt	0.01	0.01	0.00	0	0.00	−0.01	−0.04

Table 5: Gamma example, $n = 10, \theta_0 = 1$, indicating the distribution of z-value (5.22) for various non-null choices of θ. Standard deviation increases with θ_1 in accordance with (5.21), while maintaining near-perfect normality for Z_0.

So far we have considered z-values obtained from an average $\bar{y}$ of iid observations, but the results of Theorem 2 hold in greater generality. Section 5 of Efron (1987) considers one-parameter families where $\hat{\theta}$, an estimator of θ, has MLE-like asymptotic properties in terms of its bias, standard deviation, skewness and kurtosis,

$$\hat{\theta} \sim \left\{\theta + \beta_\theta/n, \sigma_\theta/\sqrt{n}, \gamma_\theta/\sqrt{n}, \delta_\theta/n\right\}. \tag{5.23}$$

Letting $\hat{\theta}$ play the role of $\bar{y}$ and $\mu_\theta = \theta + \beta_\theta/n$ in definitions (5.5)–(5.12), Lemma 3 and Theorem 2 remain true, assuming only the validity of the Cornish–Fisher transformations (5.9)–(5.12). Ignoring the bias β_θ, i.e., taking $Y_0 = \sqrt{n}(\hat{\theta} - \theta_0)/\sigma_0$ at (5.5), adds an $O(n^{-\frac{1}{2}})$ term to M in (5.17).

Moving beyond one-parameter families, suppose $\mathcal{F}$ is a p-parameter exponential family, having densities proportional to $\exp\{\eta_1 x_1 + \eta_2' x_2\} g_0(x_1, x_2)$, where η_1 and x_1 are real-valued while η_2 and x_2 are $(p-1)$-dimensional vectors, but where we are only interested in η_1, not the nuisance vector η_2. The *conditional* distribution of x_1 given x_2 is then a one-parameter exponential family with natural parameter η_1, which puts us back in the context of Theorem 2. Remark H of Section 6 suggests a further extension where the parameter of interest "θ" can be a general real-valued function of η, not just a coordinate such as η_1.

δ:	0	1	2	3	4	5
mean	0	0.98	1.89	2.71	3.41	4.01
sd	1	0.98	0.92	0.85	0.77	0.71
skew	0	−0.07	−0.11	−0.11	−0.10	−0.07
kurt	0	0.02	0.06	0.08	0.09	0.07

Table 6: Non-central t example $t \sim t_\nu(\delta)$ for $\nu = 20, \delta = 0, 1, 2, 3, 4, 5$; moment parameters of $z = \Phi^{-1}(F_\nu(t))$ (1.3) indicate near-normality even for δ far from 0. (Moments calculated using (6.10).)

The non-central t family does not meet the conditions of Lemma 3 or Theorem 2: (5.1) is symmetric in δ around zero, causing γ_0 in (5.14) to equal zero and likewise the derivative in (5.20). Nevertheless, as Figure 7 shows, it does exhibit impressive non-null normality. Table 6 displays the moment parameters of $z = \Phi^{-1}(F_\nu(t))$ (1.3), for $t \sim t_\nu(\delta), \nu = 20$ and $\delta = 0, 1, 2, 3, 4, 5$. The non-null normality isn't quite as good as in the gamma example of Table 5, but is still quite satisfactory for its application in Section 4.

Microarray studies can be more elaborate than two-sample comparisons. Suppose that in addition to the $N \times n$ expression matrix $\boldsymbol{X}$ we have measured a primary response variable y_j and covariates $w_{j1}, w_{j2}, \ldots, w_{jp}$ on each of the n subjects. Given the observed expression levels $x_{i1}, x_{i2}, \ldots, x_{in}$ for gene i, we could calculate t_i, the usual t-value for y_j as a function of x_{ij}, in a linear model that includes the p covariates. Then

$$z_i = \Phi^{-1}\left(F_{n-p-1}(t_i)\right) \tag{5.24}$$

is a z-value (1.1) under the usual Gaussian assumption, showing behavior like that in Table 6 for non-null genes.

6 Remarks

Some remarks, proofs, and details relating to the previous sections are presented here.

A. Poisson regression The curve $\hat{f}(z)$ in Figure 1 is a Poisson regression fit to the counts y_k, as a natural spline function of the bin centers x_k. Here the x_k ranged from -7.8 to 7.8 in steps of $\Delta = .2$, while the spline had five degrees of freedom, so $\boldsymbol{M}$ in (4.11) was 79×6 (including the intercept column). See Section 5 of Efron (2007b).

B. Table 3 Section 3 of Efron (2008) defines the *non-null counts* $y_k^{(1)} = (1 - \widehat{\text{fdr}}(x_k)) \cdot y_k$. Since, under model (4.2), $1 - \widehat{\text{fdr}}(x_k)$ estimates the proportion of non-null z-values in bin k, $y_k^{(1)}$ estimates the number of non-nulls. The $y_k^{(1)}$ values are plotted below the x axis in Figure 1, determining the "left" and "right" distribution parameters in Table 3. "Center" was determined by the empirical null fit from `locfdr`, using the MLE method described in Section 4 of Efron (2007b). This method tends to underestimate the non-null counts near $z = 0$, and also the σ_c values for the left and right classes, but increasing then to 1.68 had little effect on the dashed curve in Figure 4.

C. Proof of Lemma 1 Let $I_k(i)$ denote the indicator function of the event $z_i \in \mathcal{Z}_k$ (2.2) so that the number of z_i's from class $\mathcal{C}_c$ in $\mathcal{Z}_k$ is

$$y_{kc} = \sum_{\boldsymbol{c}} I_k(i), \tag{6.1}$$

the boldface subscript indicating summation over the members of $\mathcal{C}_c$. We first compute $E\{y_{kc}y_{ld}\}$ for bins k and l, $k \neq l$, and classes c and d,

$$E\{y_{kc}y_{ld}\} = E\left\{\sum_{\boldsymbol{c}}\sum_{\boldsymbol{d}} I_k(i)I_l(j)\right\} = \Delta^2 \sum_{\boldsymbol{c}}\sum_{\boldsymbol{d}} \varphi_{\rho_{ij}}(x_{kc}, x_{ld})(1 - \chi_{ij})/\sigma_c\sigma_d \tag{6.2}$$

following notation (2.5)–(2.10), with χ_{ij} the indicator function of event $i = j$ (which can only occur if $c = d$). This reduces to

$$E\{y_{kc}y_{ld}\} = N^2\Delta^2 p_c(p_d - \chi_{cd}/N)\int_{-1}^{1} \varphi_\rho(x_{kc}, x_{ld})g(\rho)d\rho/\sigma_c\sigma_d \tag{6.3}$$

under the assumption that the same correlation distribution $g(\rho)$ applies across all class combinations. Since $y_k = \sum_c y_{kc}$ (2.3), we obtain

$$E\{y_k y_l\} = N^2\Delta^2 \sum_c\sum_d p_c(p_d - \chi_{cd}/N)\int_{-1}^{1} \varphi_\rho(x_{kc}, x_{ld})g(\rho)d\rho/\sigma_c\sigma_d, \tag{6.4}$$

the non-bold subscripts indicating summation over classes.

Subtracting

$$E\{y_k\}E\{y_l\} = N^2\Delta^2 \sum_c\sum_d \varphi(x_{kc})\varphi(x_{ld})/\sigma_c\sigma_d \tag{6.5}$$

from (6.4) results, after some rearrangement, in

$$\begin{aligned}\text{cov}(y_k, y_l) = N^2\Delta^2 \sum_c\sum_d \frac{\varphi(x_{kc})\varphi(x_{ld})}{\sigma_c\sigma_d}\left\{p_c\left(p_d - \frac{\chi_{cd}}{N}\right)\int_{-1}^{1}\left(\frac{\varphi_\rho(x_{kc}, x_{ld})}{\varphi(x_{kc})\varphi(x_{ld})} - 1\right)g(\rho)d\rho\right\} \\ - N\Delta^2 \sum_c p_c \frac{\varphi(x_{kc})\varphi(x_{ld})}{\sigma_c\sigma_d}.\end{aligned} \tag{6.6}$$

Using $\pi_{kc} = \Delta \cdot \varphi(x_{kc})/\sigma_c$ as in (2.8), expression (6.6) is seen to equal the klth element of $\mathbf{cov}(\boldsymbol{y})$ in Lemma 1, when $k \neq l$.

The case $k = l$ proceeds in the same way, the only difference being that $N\Delta p_c \chi_{cd} \varphi(x_{kc})/\sigma_c$ must be added to formula (6.3). This adds $N\Delta \sum_c p_c \varphi(x_{kc})/\sigma_c$ to (6.6), again in agreement with $\mathbf{cov}(\boldsymbol{y})$ in Lemma 1.

The assumption that $g(\rho)$ is the same across all classes can be weakened for the rms approximations (2.23) and (2.33), where we only need the second moments α_2 to be the same. In fact, the class structure can disappear entirely for rms formulas, as seen in (3.12).

D. Model (2.24) Specifications (2.24) were recentered to give overall expectation 0 in (2.4), (2.5):

$$(p_0, \mu_0, \sigma_0) = (.95, -.125, 1) \quad \text{and} \quad (p_1, \mu_1, \sigma_1) = (.05, 2.38, 1), \tag{6.7}$$

these being the parameter values used in Figures 2, 3, 5 and 6. Recentering overall expectations to zero is common in practice, a consequence of the data matrix $\boldsymbol{X}$ having its column-wise means subtracted off.

The 6000×80 data matrices $\boldsymbol{X}$ used in the simulations for Table 2 and Figure 5 had entries $x_{ij} \sim \mathcal{N}(\delta_{ij}, 1)$ independent across columns j: $\delta_{ij} = 0$ for $j \leq 40$, while for columns $j > 40$,

$$\delta_{ij} = .224\,\mu_1 \quad \text{for } i = 1, 2, \ldots, 300, \quad \text{and} \quad \delta_{ij} = .224\,\mu_0 \quad \text{for } i > 300; \tag{6.8}$$

z-values based on the difference of means between the last and first 40 "patients" then satisfy (2.4), (6.7). The correlation distribution $g(\rho)$ was supported on two points, 20% $\rho = .20$ and 80% $\rho = -.05$, giving $\alpha = .10$.

E. Leukemia data standardization The original entries x_{ij} of the leukemia data matrix were genetic expression levels obtained using Affymetrix oligonucleotide microarrays. For the analyses here, each column of $\boldsymbol{X}$ was replaced by its normal score values $\tilde{x}_{ij} = \Phi^{-1}((r_{ij} - .5)/7128)$, where r_{ij} was the rank of x_{ij} in its column. Transformations such as this reduce the disturbing effects of sensitivity differences between microarrays; see Bolstad, Irizarry, Astrand and Speed (2003).

F. A parametric bootstrap method Section 3 of Efron (2007a) discusses a hierarchical Poisson simulation scheme that can be adapted to the more general context of this paper. Following the notation in (3.11), we first simulate a vector $\boldsymbol{u}$,

$$\boldsymbol{u} = N\Delta\left(\hat{\boldsymbol{f}} + \frac{\sigma_0^2}{\sqrt{2}} A \hat{\boldsymbol{f}}^{(2)}\right) \qquad \text{with } A \sim \mathcal{N}(0, \hat{\alpha}^2), \tag{6.9}$$

and then take $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{u})$, that is $y_k \overset{\text{ind}}{\sim} \text{Poisson}(u_k)$. The simulated $\boldsymbol{y}$ vectors can then be used to assess the variability of any function $Q(\boldsymbol{y})$, obviating the need for the derivative matrix $\hat{\boldsymbol{D}}$. This amounts to a parametric bootstrap approach to accuracy estimation. It produced similar answers to (4.11) when applied to the leukemia data but seemed prone to biases in other applications. (Nonparametric bootstrapping, resampling columns of the data matrix $\boldsymbol{X}$, can produce erratic results for the kind of large-scale accuracy problems considered in Section 4.)

G. z-value densities Suppose test statistic t has possible densities $\{f_\theta(t), \theta \in \Theta\}$, with corresponding cdfs $F_\theta(t)$, and we wish to test $H_0 : \theta = \theta_0$. The z-value statistic $z = \Phi^{-1}\{F_{\theta_0}(t)\}$ then has densities

$$g_\theta(z) = \varphi(z) f_\theta(t)/f_{\theta_0}(t). \tag{6.10}$$

The density curves in Figure 7 were obtained from (6.10), with $f_\theta(t)$ the noncentral $t_\nu(\theta)$ density, $\nu = 20$ and $\theta = 0, 1, 2, 3, 4, 5$.

H. Extensions of Theorem 2 In some circumstances, Theorem 2 can be extended to multiparameter families $\mathcal{F} = \{F_\eta\}$ where we wish to test $\theta = \theta_0$ for θ a real-valued function of η. This is straightforward to verify in the context of Efron (1985), which includes for example Fieller's problem, and is conjectured to be true in general exponential families.

7 Summary

The paper considers studies where a large number N of cases are under investigation, N perhaps in the hundreds or thousands, each represented by its own z-value z_i, and where there is the possibility of substantial correlation among the z_i's. Our main result is a simple approximation formula for the accuracy of summary statistics such as the empirical cdf of the z-values or an estimated false discovery rate. The argument proceeds in five steps:

- Exact formulas for the accuracy of correlated z-value cdfs are derived under normal distribution assumptions (Section 2).
- Simple approximations to the exact formulas are developed in terms of the root mean square correlation of all $N \cdot (N-1)/2$ cases (Section 2, (2.23) and (2.33)).
- Practical estimates for the approximation formulas are derived and demonstrated through simulations and application to a microarray study (Section 3 and Section 4).
- Delta-method arguments are used to extend the cdf results to more general summary statistics (Section 3 and Section 4).
- Under reasonable assumptions, it is shown that z scores tend to have nearly normal distributions, even in non-null situations (Section 5), justifying application of the theory to studies in which the individual variates are z-values.

Our main conclusion is that by dealing with normal variates, a practical assessment of large-scale correlation effects on statistical estimates is possible.

References

Bolstad, B., Irizarry, R., Astrand, M. and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193.

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* 37: 332–358.

Csörgő, S. and Mielniczuk, J. (1996). The empirical process of a short-range dependent stationary sequence under Gaussian subordination. *Probab. Theory Related Fields* 104: 15–25.

Desai, K., Deller, J. and McCormick, J. (2009). The distribution of number of false discoveries for highly correlated null hypotheses. *Ann. Appl. Statist.* Submitted, under review.

Dudoit, S., Laan, M. J. van der and Pollard, K. S. (2004). Multiple testing. I. Single-step procedures for control of general type I error rates. *Stat. Appl. Genet. Mol. Biol.* 3: Art. 13, 71 pp. (electronic).

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* 18: 71–103.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* 72: 45–58.

Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* 82: 171–200, with comments and a rejoinder by the author.

Efron, B. (2007a). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* 102: 93–103.

Efron, B. (2007b). Size, power and false discovery rates. *Ann. Statist.* 35: 1351–1377.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* 23: 1–22.

Efron, B. (2009). Are a set of microarrays independent of each other? *Ann. Appl. Statist.* To appear.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286: 531–537.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. New York: Springer-Verlag.

Johnson, N. L. and Kotz, S. (1970). *Distributions in statistics. Continuous univariate distributions. 1*. Boston, Mass.: Houghton Mifflin Co.

Lancaster, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* 29: 719–736.

Owen, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67: 411–426.

Qiu, X., Brooks, A., Klebanov, L. and Yakovlev, A. (2005a). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* 6: 120.

Qiu, X., Klebanov, L. and Yakovlev, A. (2005b). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.* 4: Art. 34, 32 pp. (electronic).

Schwartzman, A. and Lin, X. (2009). The effect of correlation in false discovery rate estimation. Biostatistics Working Paper Series number 106, Harvard University.

Westfall, P. and Young, S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York, NY: Wiley-Interscience.